Psychological Monographs

General and Applied

No. 406
1955

N. L. Gage, George S. Leavitt, and George C. Stone

Vol. 69
No. 21

Teachers' Understanding of Their Pupils and Pupils' Ratings of Their Teachers

By

N. L. Gage, George S. Leavitt, and George C. Stone

Bureau of Educational Research
University of Illinois

Price $1.50

Edited by Herbert S. Conrad

Published by The American Psychological Association, Inc.

Psychological Monographs: General and Applied

Combining the *Applied Psychology Monographs* and the *Archives of Psychology* with the *Psychological Monographs*

MANUSCRIPTS should be sent to the Editor.

Because of lack of space, the *Psychological Monographs* can print only the original or advanced contributions of the author. *Background and bibliographic materials must, in general, be largely excluded* or kept to an irreducible minimum. Statistical tables should be used to present only the most important of the statistical data or evidences.

The first page of the manuscript should contain the title of the paper, the author's name, and his institutional connection (or his city of residence). Acknowledgments should be kept brief, and appear as a *footnote* on the first page. No table of contents need be included. For other directions or suggestions on the preparation of manuscripts, see Cowan, H. S. Preparation of manuscripts for publication as monographs. *J. Psychol.*, 1948, [illegible]

CORRESPONDENCE CONCERNING BUSINESS MATTERS (subscriptions, sales, missing issues and notice of change of address, etc.) should be addressed to the American Psychological Association, Inc., 1515 Thirteenth St. N.W., Washington 6, D.C. Address changes must arrive by the 15th of the month to take effect the following month. Undelivered copies resulting from address changes will not be replaced; subscribers should notify the post office that they will guarantee third-class forwarding postage.

PUBLISHED, 1956, by the AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.

Vol. 69, No. 21 Whole No. 406, 1955

Psychological Monographs: General and Applied

Teachers' Understanding of Their Pupils and Pupils' Ratings of Their Teachers[1]

N. L. Gage, George S. Leavitt, and George C. Stone
Bureau of Educational Research
University of Illinois

I. PURPOSE AND SCOPE OF THE STUDY

THE PROPOSITION that teachers should understand their pupils receives well-nigh universal assent. Because of the sheer reasonableness of this idea, many of its assumptions have remained unanalyzed. We have undertaken research on this proposition to contribute both to theory of the teaching process and to practice in teacher selection and training.

Analyzing this proposition means delimiting its terms and specifying their dimensions, both conceptually and operationally. By *teacher*, thus, do we mean anyone who influences any kind of learning on the part of any other person in any kind of situation? For example, do we mean nursery school teachers *and* graduate school professors, psychotherapists *and* army officers, parents *and* salesmen? Our basic statement would probably become more meaningful if we specified the level of pupil maturity, the allowable permissiveness, and the intimacy of the teacher-pupil relationship, to mention only a few of the possible dimensions of the teacher's role. In this study, therefore, we concentrate on one kind of teacher.

Similarly, should we refine what we mean by *understand*? Does this term refer to qualitative or quantitative judgments about others? Does it refer to easily expressed, rationalized, intellectual processes, or to ineffable, emotionalized intuitions? Does it refer to knowledge of directly observed facts about the other or to inferences from cues that have no obvious relevance? In this study we deal with the teachers' understanding as evidenced by their ability to predict three kinds of pupil behavior. (We use the terms "understanding" and "perception" interchangeably; this is in accord with much current usage.)

Thirdly, what do we mean by *pupil*? What characteristics or behaviors of the pupil do we have in mind? Do we mean understanding of the typical pupil or the individual pupil, of something as specific as the pupil's IQ or something as general as his "personality?" Answers here depend, of course, on many considerations. In this study, three kinds of pupil behavior were chosen according to considerations we shall seek to make explicit.

In short, this investigation was aimed at formulating and empirically validating hypotheses stemming from analysis of the basic proposition that teachers should understand their pupils. By

[1] This investigation was supported by a research grant (M-650) from the Institute of Mental Health, National Institutes of Health, Public Health Service.

selecting, from the many possibilities, a delimited set of specific instances of (*a*) teachers, (*b*) pupils, and (*c*) understandings, we have attempted to test some of the relationships that would justify the hortative *should*.

A. The Teachers and Their Pupils

We studied all 103 fourth-, fifth-, and sixth-grade teachers and their 2,885 pupils in the 19 elementary schools of a Midwestern city of about 65,000 population. Eighty-four of the teachers were women. Only seven classes combined two grades. Class size ranged from 19 to 40 with a mean of 28.

These grade levels were chosen because (*a*) their pupils have only one teacher all day for an entire school year, so that the teacher-pupil relationship is presumably more salient than at the junior and senior high school levels, (*b*) the pupils in these grades can read and write well enough to make paper-and-pencil testing workable, and (*c*) by using three grade levels we could not only obtain a larger sample of teachers but study pupil maturity, over at least a narrow range, as a variable affecting the relationships between other variables. The city was chosen for its size and hence for its provision of an adequate sample of teachers within a single school system, and of course for its reasonable proximity to our own offices.

The teachers and pupils cooperated with the understanding that their scores and other data would be kept confidential and never be allowed to influence anyone's professional or personal standing. The data were collected by graduate students who were given detailed instructions and training. Teachers took their tests outside the classroom while the pupils were filling out their forms and inventories inside the room. In all, we used about one hour of concurrent testing time for both teachers and pupils.

B. Three Classes of Variables

Whether teachers should understand their pupils depends on whether there are positive relationships between teachers' understanding of their pupils and certain other phenomena considered desirable. We are concerned from the start, therefore, with two kinds of variables: *teachers' understanding*, and valued phenomena. In this study the valued phenomena we have chosen for investigation are *pupils' favorable descriptions of their teachers*. That is, we have more or less arbitrarily decided that pupils' favorable descriptions of their teachers can justify the value of teachers' understandings of pupils. This is not the place for any extended consideration of the significance of pupils' attitudes toward or beliefs concerning their teachers. Our position is simply that such attitudes and beliefs *are* educationally significant, quite apart from their relationship to such variables as objectively measured achievement. Further, such ratings are useful in research because pupils have greater opportunity than anyone else to observe the teacher, and they are sufficient in number to make their pooled ratings highly reliable.

In the course of measuring teachers' understandings, we had to generate *data concerning the pupils*. These data provided the veridicalities against which we have measured the accuracy of teachers' perceptions of pupils. Since we attempted to choose educationally significant areas of pupil behavior in which to measure teachers' understandings, the data on pupils are worth analyzing in

their own right. Results of such analyses will be presented in a subsequent report.

The succeeding sections of this report present (*a*) our measures of teachers' understandings of pupils (Section II), (*b*) pupils' descriptions of their teachers (Section III), (*c*) the relationships between them (Section IV), the implications of the investigation (Section V), and a summary (Section VI).

II. TEACHERS' PERCEPTIONS OF PUPILS' CHARACTERISTICS

The kinds of perceptions investigated in this study are named according to the characteristics or behaviors of pupils whom the teachers were asked to perceive. Largely on a priori grounds, three such aspects of pupils were identified: cognitive, social, and emotional. Definitions of these are presented below, both conceptually and operationally. In each case, we developed a test of teachers' accuracy in perceiving the given aspect of pupils.

A. Measuring Teachers' Accuracy in Perceiving Cognitive Aspects of Pupils

Cognitive aspects of pupils are defined as their knowledges and their intellectual abilities and skills. Other terms for what we here have in mind are "intelligence," "mental ability," and "educational achievement." These aspects of pupils are prime determiners of ability to achieve school objectives in the traditional sense. Whether determined by heredity, by home environment and background, or by previous school influences, these cognitive aspects of a pupil affect his understanding of a text book, of a teacher's explanation, or of many problems arising in school work.

To foster achievement of the cognitive objectives of education is a major part of the teacher's role. In doing this, teachers must begin "where the pupils are." That is, teachers must judge the intellectual readiness of pupils—their ability to learn from a task, to understand an explanation, to solve a problem. Conceivably, teachers who make such judgments more accurately will, other things equal, teach more effectively. Their assignments, their explanations, and their discussions will be more appropriate to their pupils' abilities.

Possible Operational Definitions

How can the teacher's accuracy in perceiving cognitive aspects of pupils be measured? Among the possibilities that immediately suggest themselves are the following:

Ask teachers to estimate pupils' IQ's, and score the estimates against actual IQ's as determined by intelligence testing.

Ask teachers to predict pupils' answers to intellectual problems of either the free-choice or multiple-choice type, and score the predictions against the pupils' actual answers.

Ask teachers to predict the relative popularity among their pupils of the four or so response alternatives in multiple-choice intelligence- or achievement-test items, and score the predictions against the actual rank order in popularity of the various alternative responses.

Ask teachers to estimate the percentage of their pupils who will respond correctly to a series of intelligence- and achievement-test items and score the predictions against the actual percentages.

Each of these procedures would have required collecting data on the teacher's own pupils against which to score her estimates or predictions. Because of limited time for testing pupils in this investigation, and because available time was needed for other purposes, none of the above possibilities was pursued.

Development of the "Which Question Is Harder?" Test

Instead, a fifth possibility was developed. This was to ask the teacher to estimate the relative difficulty (as against the absolute difficulty, which we thought

might allow irrelevant response sets to appear) of multiple-choice vocabulary and arithmetic items taken from a nationally standardized achievement test designed for grades four through six.[3] Data on the percentage passing each item in the nationwide standardization sample of pupils were obtained from the publisher of the test. The three percentages of pupils in Grades 4, 5, and 6 who passed each item were averaged to obtain a single difficulty index for each item. These indices were used to make the key with which teachers' rankings of the items were scored for accuracy.

Three forms of this test were tried. Form 1 contained four sets of ten multiple-choice items each. Two of the sets contained vocabulary items, and two contained arithmetic items. The items in each set were chosen so as to represent, in roughly equal steps, all levels of difficulty from about 90 per cent passing to about 10 per cent passing. The test was administered to 56 undergraduate elementary education majors, with the following directions:

"This is a test of your ability to judge the relative difficulty of test questions. You are given sets of ten questions from tests for Grades 4-6. In each set, you are to rank the items as to their difficulty.

"Your rankings will be scored against the actual rankings of the questions, as determined by the percentages of pupils who answered each item correctly. These difficulty percentages were determined by the test publishers after testing a nationwide sample of pupils in Grades 4-6.

"You should read the directions given to the pupils, the problem itself, and the choices from which the pupil had to select his answer. Remember that the difficulty depends on the choices as well as on the problem.

"First, decide which problem is *easiest*, and write the number '1' before that problem. Then decide which problem is *hardest*, and write '10' before that problem. Write '2' before the second easiest and '9' before the third easiest, etc., until you have ranked all ten problems in the set.

"Obviously, you will need to read all ten problems in a set before you assign difficulty ranks."

The reliability of this test, estimated by correlating rank-difference squared scores on one half against those on the other half, proved to be only .16, corrected by the Spearman-Brown formula.

Consequently, a second form was developed. We suspected that the first form was unreliable because each ten-item ranking task elicited in effect only a relatively small number of independent judgments. Hence, the second form was designed to require more such independent judgments. Form 2 contained 60 *pairs* of achievement-test items to be judged according to the following directions:

"This is a test of your ability to judge the relative difficulty of test questions. You are given pairs of questions from tests for Grades 4-6. In each pair you are to decide which you think is the harder.

"Your decisions will be scored against the actual rankings of the questions, as determined by the percentages of pupils who answered each item correctly. These difficulty percentages were determined by the test publishers after testing a nationwide sample of pupils in Grades 4-6.

"You should read the directions given to the pupils, the problem itself, and the choices from which the pupil had to select his answer. Remember that the difficulty depends on the choices as well as on the problem.

"Decide which problem is *harder* and make a check in the blank opposite that problem.

"Every pair of questions differed by at least 20 per cent in the percentage of pupils answering correctly."

The items were paired so that the "difficulty gap" (difference between the percentages passing the two items) ranged from about 70 per cent to about 20 per cent, with roughly equal numbers of pairs at each point of the difficulty gap continuum. Data obtained by administering this form to 73 undergraduate elementary education majors yielded a Spearman-Brown corrected coefficient of internal consistency of .34.

It seemed possible that the reliability of this test could be improved by adjusting the difficulty gap of the 60 pairs of items presented. To do this, the difficulty gap of each pair was plotted against the percentage of correct judgments of that pair by the judges who took Form 2. As expected, a marked positive relationship appeared. Furthermore, the plot indicated that the difficulty gap should be about 35 per cent to obtain the 75 per cent of correct judg-

[3] This was the Stanford Achievement Test, published by World Book Company, to whom we are grateful for the data on item difficulty and for permission to use the items in our research.

TABLE 1
THE "WHICH QUESTION IS HARDER?" TEST, FORM 3

Group	*N*	M	*SD*	r_{11}*	$r_{bis.}$
Undergraduates in elementary education	103	46.64	4.16	.45	.32
Teachers of fourth, fifth, and sixth grades	103	48.67	4.10	.52	

* As estimated with Kuder-Richardson formula 20.

ments of each pair that would maximize efficiency of measurement. Accordingly, Form 3 was constructed. The 60 pairs of items in Form 3 had difficulty gaps ranging from 23 per cent to 47 per cent, with a median of about 28 per cent. Directions for Form 3 were identical with those for Form 2. Administered to 103 undergraduate elementary education majors, it yielded a reliability coefficient of .45 (Kuder-Richardson formula 20).

Although low in internal consistency, so that much error variance was obtained with the test, Form 3 seemed to have enough logical validity to be worth administering for exploratory purposes. Table 1 shows the results obtained from giving the test both to the tryout sample of 103 teachers-in-training and to the 103 teachers-in-service. The difference between the means of the two groups is statistically significant beyond the .01 level. The biserial *r* between Form 3 and teaching experience is .32, also significant beyond the .01 level. Experienced teachers presumably have had more experience than teachers-in-training with the cognitive aspects of pupils at which the test was aimed. Hence the success of the test in distinguishing between the two groups indicates positive empirical validity. Further data on the correlates of this test are given in subsequent sections. (See page 33 for a discussion of the possible consequences of our procedure of scoring the test against national norms, and suggested changes.)

B. MEASURING TEACHERS' ACCURACY IN PERCEIVING SOCIAL ASPECTS OF PUPILS

By social aspects of pupils we mean their "popularity," "social adjustment," "sociometric choice status," and "peer relationships" in the classroom. These aspects of pupils both determine and reflect social learnings that many American educators consider important, especially in the elementary grades. As to the reasons for attaching importance to social aspects of pupils, we shall not here be concerned. Suffice it to say that "social adjustment" is considered important not only in its own right but because its influences cognitive and emotional development.

Hence teachers in the elementary grades are usually expected to foster the social adjustment of all their pupils. Sometimes, this means preventing any pupil from becoming or remaining an "isolate," unchosen or unwanted by any of his classmates. It may mean that no pupil should become too much a "star," so that life of the classroom is unhealthily dominated by his values and other pupils are deprived of opportunities to lead.

To promote a desirable sociometric structure in the classroom, the teacher should presumably understand what the current structure is. Conceivably, teachers who perceive these social aspects of their pupils more accurately will, other things equal, be more effective in pro-

moting healthy social relationships among their pupils.

A Test of the Teacher's Ability to Judge Interpupil Preferences

Various methods for the measurement of accuracy in perceiving social relationships of pupils have already been used (1, 7, 9). These procedures either did not apply or could be improved for our purposes. Dymond's (7) method dealt with accuracy in perceiving others' attitudes towards *oneself*. Ausubel's (1) and Gronlund's (9) methods applied to accuracy in judging the *over-all* popularity of a pupil. But a teacher might conceivably be accurate in this sense while having in mind the wrong friends for any given pupil. It seemed preferable to measure the teacher's accuracy in predicting the *specific* classmates that each pupil would choose. Such a procedure would at the same time yield data for scoring accuracy in judging over-all sociometric status. Accordingly, the 103 teachers in our study were given the following directions:

"This is a test of your ability to judge the preferences of pupils for other pupils. The pupils in your class are being asked to name the five classmates 'whom they would most like to have in the same class with them if this class were to be divided into two groups.'

"To what extent can you 'predict' whom each pupil will choose? Sometimes several pupils will choose the same pupil as their preferred friend. Others may choose pupils no one else chooses. Also, as you know, children at the fourth-, fifth-, and sixth-grade levels seldom choose members of the opposite sex as friends.

"Your predictions of each pupil's preferences will be scored for accuracy against the actual preferences of that pupil.

"1. On the following page, please write the names of your pupils as they are in your class book. (Place last name first.)

"2. After the name of each pupil, write the number of the two pupils that you think he or she will choose as his or her preferred companions."

Example:

	Nos. of 2 Pupils Chosen	
1. Adams, Joe	*14*	*3*
2. Baker, Sue	*27*	*26*
3. Carr, Bill	*9*	*14*
:		
26. Williams, Mary	*13*	*27*
27. Young, Ann	*26*	*19*
28. Zenger, Al	*8*	*9*

The directions to pupils for choosing their preferred companions were as follows:

"Suppose that your class were to be divided into two groups with different teachers in different rooms. Write the names of the five pupils whom you would most like to have in the same group with you. Choose anyone in this room you wish. You may choose pupils who are absent today. Your choices will not be mentioned to anyone else. Give both first and last names. Spell them the best you can. Write the name of your first choice after Number 1, your second choice after Number 2, and so on.

"I would like to have these children in the same group with me:

1. _____
2. _____
3. _____
4. _____
5. _____."

Scoring the test for "specific" accuracy presented some problems. It was easy to give each teacher one point for each correct prediction as to a pupil's choice of a classmate. But teachers with larger classes would have larger scores, because they made more predictions. So we divided each teacher's number-right score by the number of pupils for whom the teacher made predictions. But the probability of chance accuracy is less in larger classes, since the teacher has more pupils to choose from in predicting any pupil's

choices. Thus larger class size operated both to raise and lower scores. To correct for the latter kind of influence, the ratio (number right/number of pupils) scores were correlated with class size. The r of $-.35$ was used to obtain a regression-on-class-size which was applied to the obtained score so as to yield what we called the *regression-corrected ratio score.*

The two corrections entering into this score take into account mathematical effects of class size on the accuracy score. The regression correction may also, however, take out variance due to valid *psychological* effects of class size. That is, if it *is* more difficult to perceive pupil's preferences for one another in a larger class, simply because there are more pupils to keep track of, this probably should be allowed to influence the score. Our regression correction may remove such influences along with the mathematical ones, and to that extent might make the resulting score less logically valid. But evidence from our other score, presented below, indicates that this criticism is unwarranted.

The regression-corrected ratio scores of the 103 teachers had a mean of 1.25 and a standard deviation of .20. The reliability of the test, estimated by using Guttman's L_4 formula (12) on the regression-corrected ratio scores obtained from the odd-numbered ten and even-numbered ten of the first 20 pupils in each class, was .52. This value is, of course, an underestimate, since 98 of the 103 classes had more than 20 pupils. Corrected to the mean class size, 28 pupils, this reliability becomes .60.

Scoring the test for over-all correlational accuracy was relatively straightforward. We counted the number of choices the teacher predicted each pupil would receive. This number was correlated against the number each pupil actually did receive. The resulting r was the teacher's *correlational accuracy score.*[3] For 103 teachers these correlational scores had a mean of .48 and a standard deviation of .19. The mean shows that the teachers had considerably better-than-chance success in judging the relative sociometric status of their pupils; it is similar to the mean correlational score of .595 achieved by the teachers in Gronlund's investigation (9, p. 35).

How are the correlational accuracy scores related to class size? (The regression-corrected ratio score is by definition, of course, unrelated to class size.) The r between correlational score and class size is .10, essentially the same as the value of $-.007$ reported by Gronlund (9, p. 39). This means that there is probably no psychological effect of class size on accuracy in perceiving sociometric aspects of pupils. At least within the ranges of class size in our sample (19 to 40 pupils) and in Gronlund's (15 to 43 pupils), teachers are not handicapped in this task by having more pupils to keep track of. Hence it is unlikely that, as we supposed above, our regression-corrected ratio score has had some of its psychological significance "corrected out" along with the mathematical effect of class size.

The reliability of the correlational accuracy score was not estimated because there seemed to be no acceptable method for our data. Neither Kuder-Richardson nor split-half methods could be applied, for rather obvious reasons that need not be detailed here.

The regression-corrected ratio score correlated .46 with the correlational

[3] We have used this score without Fisher's z transformation. For r's of the size obtained, the z transformation would not change the results appreciably. (See 10, pp. 355-356.)

score. Although positive, as expected, this r indicated that the two scoring methods were measuring different kinds of accuracy in perceiving social aspects of pupils.

C. Measuring Teachers' Accuracy in Perceiving Pupils' Problems

By the problems of pupils we mean the tensions, frustrations, and unmet needs symptomatic of the adjustment process. Problems can be so profound that repression has driven them out of awareness. Or we can refer to problems in their more superficial guises—the worries, bothers, and wishes that all of us carry to the surface. Whether the two levels are linked in any systematic way is a problem that need not concern us here. We need only assume that problems which pupils are able and willing to identify in themselves can have some significance for the educational process.

Teachers nowadays are trained to understand that "the whole pupil comes to school, not merely the to-be-taught-subject-matter pupil." This point of view arises from two related doctrines of modern education. (*a*) The teacher, especially in the elementary grades, has responsibilities for more than the intellectual development of pupils; mental health must also be the teacher's concern. This is because the teacher and the school situation will inevitably affect the mental health of pupils, one way or another, regardless of whether or not the teacher pays attention to this facet of her influence. (*b*) Even if the teacher wanted to concern herself only with cognitive objectives, the pupil's emotional life would still be important to her. This is so because school learning is influenced by the feelings, attitudes, values—indeed, problems and worries—that the pupil brings from home and neighborhood, or finds in the schoolyard and classroom.

So it has come to be believed that the teacher can best foster both the intellectual growth and the mental health of her pupils if she "knows what is on their minds." The girl who is worried about her complexion will be better taught by a teacher who is not oblivious of this salient fact about the pupil's "inner life." The boy who feels he has no friends will be handled better by the teacher who knows that this is what is "eating" him, during the arithmetic lesson as well as on the playground.

These, at least, are assumptions behind much teacher-training in present-day America, behind the study of mental hygiene required of many prospective teachers. One purpose of mental hygiene courses, often stated, is to increase the sensitivity of teachers to the emotional life of pupils. Greater awareness and understanding should, other things being equal, result in more appropriate behavior of teachers when they impinge upon the emotional security of their pupils.

It was realized that in distinguishing between social and emotional aspects of pupils—and between the corresponding kinds of perceptual accuracy and behavior on the part of teachers—we were probably being more logical than psychological. However logically distinct these two aspects might be, we expected to find them substantially related empirically. That is, pupils' social aspects are often emotional aspects also, and vice versa. The distinction between these two aspects is much less sharp than that to be made and found between them and cognitive aspects. These prior considerations will help in understanding some of the results reported below.

Possible Operational Definitions

One approach to the definition and measurement of teachers' understanding of pupils has been the opinion-attitude inventory, exemplified by the Minnesota Teacher Attitude Inventory (3) and the How I Teach questionnaire (17). These inventories consist of statements concerning pupils, teaching, and teacher-pupil relationships, on which the teacher is asked to agree or disagree. They seem to measure permissive, supportive orientation toward pupils, but they may be fakable (21). Such questionnaires do not, furthermore, test the teacher's understanding of pupils against observed facts about pupils. Rather, they measure understanding of abstractions and principles that are considered to apply to pupils and teacher-pupil relations in general. The questions have remained open whether we can measure individual

differences among teachers in how accurately they understand the personal life of their own pupils, and whether such measures correlate with the appropriateness of the teacher's behavior in meeting her pupils' needs for emotional acceptance and support.

Ideally, such a test would ask in unstructured form for the teacher's description of the pupil's emotional life from the pupil's point of view. Somehow an account of this kind, minimally distorted by the experimenter, would also be obtained from the pupil. The two would be compared and, from their similarity, the degree of the teacher's understanding would be inferred. So, if the pupil feels himself driven by the example of his younger brother and also dreads reciting in class because of a conviction that his voice is effeminate, the perceptive teacher will somehow have discerned these feelings. Perhaps she will, even if somewhat unintentionally, offer support in just these areas, whether by judicious praise, by removing pressure, or by fostering insight.

Development of the Problem Prediction Test

Such an individualized procedure may prove eventually to be the only kind that can have any validity. Yet it would be extremely costly. Hence, it seemed worth while to try an easier approach. This was to ask the teachers to predict how some of their pupils would fill out a forced-choice problem inventory. The statements of problems were taken from the SRA Junior Inventory. They were grouped into sets of three on the basis of their approximately equal prevalence among fourth- to sixth-grade pupils; prevalence was judged from the percentages, reported in the manual, of a nationwide group of such pupils who had checked the problems as "things that bother or worry us."

The forced-choice format was chosen to reduce the likelihood that response sets would occur on the part of the pupil and teacher. If the pupils had been allowed to check each statement independently of the others, their responses, as much research has shown, probably would have reflected reliable but largely irrelevant tendencies to "check when in doubt." Teachers' predictions also would have been influenced by such a response set. As has been demonstrated (23), teachers' predictive accuracy scores would then have been greatly influenced by chance agreement between their response sets and those of their pupils. Such an influence on the scores would make them less relevant to the understanding at which the test was aimed.

Accordingly, we asked all pupils to fill out the inventory by writing, in each set of three problems, the number 1 before the problem "which worries you most," the number 3 before the problem "which worries you least," and the number 2 beside the remaining one. The teachers were asked to fill out the same inventory *for eight selected pupils* according to the instructions at the top of the following page (page 10).

The test was scored by summing the squared differences between the pupil's ranking of each problem and the teacher's prediction of that ranking. Thus a low score is an accurate score; in reporting r's involving this score, we have reflected signs so that positive r's would show positive relationships between desirable "traits." Since there were 12 sets of three problems each, the teacher's score for any single pupil, or the mean score on eight pupils, could range from zero (perfect accuracy) to 96; a chance score is 48. The actual range of these means for 103 teachers was from 35.38 to 53.75, with a mean of 43.48 and a standard deviation of 3.99. Thus our teachers, on the average, predicted their pupils' responses with significantly better-than-chance accuracy. But this mean is equivalent to a rank-difference correlation, rho, of only .09. It is evident that the average level of accuracy, in predicting the rank order of three problems, is quite low; in other words, the task was perhaps too difficult.

The reliability of this test could be

THE PROBLEM PREDICTION TEST

This is a test of your ability to predict the problems which your pupils will say bother them most and least.

On the following pages are the (1) directions and (2) questions that are being given to your pupils concerning some of their problems. You are given eight sets of the questions—one set for each of the eight pupils whose answers you are to predict.

Since your predictions will be scored against the actual answers of the pupils, do not predict for any pupil who is absent today. That is, select only pupils who are present in class today.

Fill out the questionnaire as you think each child *actually did*—not *necessarily* as he *should* if he tells the truth. That is, *predict—do not describe*—the child.

You should predict the answers of the:

Two Boys You Find *Easiest* to Work With (Write their names below.)	Two Girls You Find *Easiest* to Work With (Write their names below.)
1. _____	5. _____
2. _____	6. _____
Two Boys You Find *Most Difficult* to Work With (Write their names below.)	**Two Girls You Find *Most Difficult* to Work With (Write their names below.)**
3. _____	7. _____
4. _____	8. _____

Write the name of the pupil for whom you are predicting at the top of the page on which you are making your predictions.

estimated in two different senses. Reliability over items indicates how accuracy on our set of twelve items would correlate with accuracy on another equivalent set of 12 items for any pupil or for all pupils. As shown in Table 2, the corrected split-half coefficients for scores based on each of the eight pupils ranged from zero to .63, with a median of .30. The corrected split-half coefficient for the total score on all eight pupils is .26. Reliability, or generality, over pupils indicates how accuracy on one pupil would correlate with accuracy on another, or how total accuracy on four of the pupils correlated with total ac-

TABLE 2

CORRECTED SPLIT-HALF RELIABILITIES OF THE TEACHERS' SCORES ON THE PROBLEM PREDICTION TEST ON EIGHT SELECTED PUPILS

Pupil	No. of Teachers	Odd Items		Even Items		Split-half *r*	
		M	*SD*	M	*SD*	Uncorrected	Corrected
1	102	22.10	7.06	21.49	7.24	−.027	—
2	100	21.70	6.97	20.74	7.05	.196	.33
3	101	22.05	7.39	23.22	6.65	.460	.63
4	96	21.22	7.13	22.86	7.14	.106	.19
5	102	22.95	7.03	18.22	6.99	−.0004	—
6	102	23.21	7.29	20.58	6.74	.112	.20
7	99	22.83	7.39	22.45	6.37	.225	.37
8	98	21.26	6.63	21.06	7.12	.162	.28
Total	81	174.63	20.33	169.16	21.12	.148	.26

TABLE 3

INTERCORRELATIONS AMONG EIGHT ACCURACY SCORES ON THE PROBLEM PREDICTION TEST ($N = 81$)

Item	M	*SD*	Intercorrelations						
			(2)	(3)	(4)	(5)	(6)	(7)	(8)
1. Boy *a*—Easiest to work with	43.16	10.05	.02	−.02	.08	.17	.18	−.05	.20
2. Boy *b*—Easiest to work with	42.60	10.84		.02	.04	.17	−.07	−.06	−.05
3. Boy *a*—Most difficult to work with	44.06	12.30			−.06	−.01	.09	.06	.00
4. Boy *b*—Most difficult to work with	42.93	10.14				.00	.12	−.14	.08
5. Girl *a*—Easiest to work with	41.14	10.80					.09	.06	−.15
6. Girl *b*—Easiest to work with	43.35	10.95						.05	.01
7. Girl *a*—Most difficult to work with	45.26	10.91							−.18
8. Girl *b*—Most difficult to work with	41.59	10.46							

curacy on the other four. Table 3 shows the *r*'s among the eight accuracy scores. The median of these is .01, the range extending from −.18 to .20. No generality of this score is evident, either within or between sexes, either within or between "easiest" and "most difficult to work with" children.

D. Intercorrelations Among the Three Tests of Predictive Accuracy

We have presented the rationale, conceptual and operational definitions, and descriptive statistics for each of three types of accuracy in perceiving characteristics of pupils. Table 4 shows how

TABLE 4

INTERCORRELATIONS AMONG TESTS OF ACCURACY IN PERCEIVING CHARACTERISTICS OF PUPILS ($N = 103$)

Test	M	*SD*	*r*'s		
			(2a)	(2b)	(3)
1. The "Which Question Is Harder?" Test, Form 3	48.67	4.10	−.02	−.04	−.04
2. A Test of Ability to Judge Interpupil Preferences:					
2a. Regression-corrected ratio score	1.25	.20		.46**	.33**
2b. Correlational score	.48	.19			.20*
3. The Problem Prediction Test	43.48	3.99			

* Significant at the .05 level.
** Significant at the .01 level.

TABLE 5

SEX DIFFERENCES IN TEACHERS' ACCURACY SCORES

Test	Male Teachers ($N=19$) Mean	*SD*	Female Teachers ($N=84$) Mean	*SD*	Difference in Means
"Which Question Is Harder?" Test	46.84	4.196	49.08	3.842	−2.24*
Prediction of Interpupil Preferences	1.31	.207	1.23	.196	.08
Problem Prediction	42.62	3.998	43.69	3.964	−1.07

* Significant at the .05 level.

these three measures intercorrelate. If they were corrected for attenuation, the *r*'s between Test 2 and 3 would indicate great overlap between these measures. Nonetheless, the empirical validity of all measures merits further investigation, since Test 3's low reliability makes the correction for attenuation very great. Also, if we grant these measures any degree of logical validity as tests of accuracy in perceiving pupils, it is evident that Tests 1 and 2, at least, cannot be subsumed under any single over-all construct, such as "empathic ability" or "social perceptiveness." Our findings here corroborate those of Gage and Exline (8), Crow (5), Bell (2), Taft (24), and others who have reported independence between tests that superficially appear to be measuring varieties of a single function such as "accuracy in perceiving the characteristics of other persons." One kind of "accuracy of social perception" seems to have little to do with another. What have in the past appealed to some investigators as aspects of a single, general, logically definable psychological process turn out empirically to be different, unrelated processes.

It is nonetheless noteworthy that the *r*'s between Tests 2 and 3 seem to corroborate our expectation that the logical distinction between social and emotional aspects of pupils, and the corresponding sensitivities to these aspects, would not be borne out empirically. Other findings, reported later, also bear on this point.

E. TEACHERS' ACCURACY IN RELATION TO OTHER CHARACTERISTICS

How are the three kinds of teacher understanding related to the teachers' sex, age, school grade, and attitudes towards pupils and teacher-pupil relationships (Minnesota Teacher Attitude Inventory)?

As shown in Table 5, male teachers were significantly less accurate ($p < .05$) than female teachers on the "Which Question Is Harder?" test. The differences on the other two accuracy scores were not significant.

The correlations of the three accuracy scores with the teachers' ages, which ranged from 22 to 62, were all essentially zero. This was also true of the relationships between the three accuracy scores and the teachers' scores on the Minnesota Teacher Attitude Inventory, taken a year previously by 80 of the teachers.

Fourth-grade teachers were more accurate than fifth- and sixth-grade teachers on the "Which Question Is Harder?" test, the difference between fourth- and sixth-grade teachers being significant at the .05 level (see Table 6). In contrast, teachers of fifth- and sixth-grade classes were more accurate ($p < .05$) in predicting interpupil preferences. This finding agrees with Ausubel's report (1) that

TABLE 6

GRADE DIFFERENCES IN TEACHERS' ACCURACY SCORES

Test	Fourth Grade $N=34$		Fifth Grade $N=33$		Sixth Grade $N=36$			Differences in Means		
	Mean	*SD*	Mean	*SD*	Mean	*SD*	*F*	4−5	5−6	4−6
"Which Question Is Harder?" Test	49.82	4.258	48.63	3.496	47.61	4.163	2.73			
Prediction of Interpupil Preferences	1.17	.185	1.28	.215	1.28	.182	3.55*	−.11*	.00	−.11*
Problem Prediction	44.50	3.570	43.16	4.172	42.84	4.011	1.74			

* Significant at the .05 level.

teachers' "socioempathy" increased with the age of the children whose sociometric choices were predicted. This increased accuracy is apparently recognized by the children, as will be seen later (p. 26).

III. TEACHERS' BEHAVIORS AS DESCRIBED BY PUPILS

How teachers perceive pupils can be considered important in itself. Systematic study of such perceptions can easily absorb much investigative effort. So far this report has described only our attempts to conceptualize and measure phenomena of this kind.

But most educators would make practical demands of any study of how teachers perceive pupils. They would concur with Sears (22, p. 477):

> There are many kinds of observations that can be and have been made of social behavior. Some of these have involved inferred traits and needs; others have related to perceptions or to states of consciousness. By the criterion of logic, a theory that takes any of these phenomena as its basic reference events is acceptable. But there is another criterion to be considered, the practical one. It is reasonable to ask what kinds of events are important to us. On this score, action is clearly more significant than perception or traits.

In short, for application to the training, selection, and supervision of teachers, we should ascertain relationships between teachers' social perceptions and their behaviors.

In particular, how do teachers' perceptions of pupils correlate with teachers' behaviors vis-à-vis pupils? To answer this question requires descriptions of such behaviors. Such descriptions must be based on observations. Three kinds of persons can observe teacher behavior: the teacher herself; outsiders, such as principals, supervisors, or specially trained observers; and pupils. Observations and descriptions by outsiders are difficult and expensive to obtain; such observers must be given special training and need considerable time for observation if they are to achieve satisfactory interobserver reliability. Pupils, on the other hand, see the teacher 25 hours a week under normal conditions, and their observations and descriptions, regardless of their "objective" validity, have intrinsic educational significance. When the ratings of 20 or so pupils are averaged, the reliability of the mean is usually found to be quite high. So have we rationalized our reliance in this study on pupils' observations and descriptions of their teachers.

A. THREE KINDS OF TEACHER BEHAVIOR

For the aspects of their teachers that we should ask pupils to describe, we turned to our categories of teachers' perceptions. For each of the three aspects

OUR TEACHER

Here are some questions about your teacher.

Read each question and decide how often your teacher does what is asked about. Underline the word that shows how often she does it.

Please answer the questions honestly. You are *not* asked to write your name on the paper.

None of the teachers or the principal will ever see this paper or know how you answered the questions. In answering the questions, think of the teacher whose name is below:

Name of Teacher: _____

Then underline the word *always* or *usually* or *sometimes* or *never* which shows what you think.

1. Does your teacher explain school work so that you can understand it?
   a) always b) usually c) sometimes d) never

2. Does your teacher know which pupils you like best in this class?
   a) always b) usually c) sometimes d) never

3. Does your teacher make you feel that she likes you?
   a) always b) usually c) sometimes d) never

4. Does your teacher make you want to learn new things?
   a) always b) usually c) sometimes d) never

5. Does your teacher make sure that no pupils get left out of things?
   a) always b) usually c) sometimes d) never

6. Does your teacher make sure not to hurt your feelings or make you feel afraid?
   a) always b) usually c) sometimes d) never

7. Does your teacher catch on quickly to what mixes you up in school work?
   a) always b) usually c) sometimes d) never

8. Does your teacher help all pupils show what they are good at?
   a) always b) usually c) sometimes d) never

9. Does your teacher know when you are trying hard?
   a) always b) usually c) sometimes d) never

10. Does your teacher make school work hard enough but not too hard?
    a) always b) usually c) sometimes d) never

11. Does your teacher see that pupils do not look foolish to the rest of the class?
    a) always b) usually c) sometimes d) never

12. Does your teacher know what worries or bothers you?
    a) always b) usually c) sometimes d) never

FIG. 1. The Unforced Teacher Rating Scale.

of pupils that teachers were asked to judge, we proposed to have the pupils rate a corresponding kind of teacher behavior. In each case, this behavior was considered from the standpoint of its *appropriateness*. That is, we defined the teacher behavior relevant to each kind of perception in terms of its appropriateness to the pupils' abilities, standings, or needs as revealed by that kind of perception. So we distinguished appropriateness of teacher behavior in promoting *cognitive, social,* and *emotional* adjustment. The concrete meaning of each of the dimensions is shown by the rating scale used to obtain these descriptions (see Fig. 1).

The "Cognitive Appropriateness" Rating-Scale Items

Items 1, 4, 7, and 10 of the "Our Teacher" scale were intended to elicit pupils' appraisals of the degree to which teachers met pupils' needs for cognitive understanding, motivation, clarification, and adjustment. The key words in these items are "explain" and "understand" (Item 1), "learn" (Item 4), "catch on" and "school work" (Items 7 and 10). They are intended to deal with the teacher's effectiveness in the traditional tasks of conveying knowledge and imparting understanding.

While all these items were felt to be relevant to the teacher's accuracy in perceiving the relative difficulty of intellectual tasks, Item 10 ("make school work hard enough but not too hard") was designed to be particularly relevant to such accuracy. Similarly, Item 1 ("explains school work so that you can understand it") should especially characterize teachers who could tell which question is harder; such a teacher would choose words and illustrations that would be better suited to pupils' capabilities.

The "Social Effectiveness" Rating-Scale Items

These were Items 2, 5, 8, and 11. Item 2 ("knows which pupils you like best in this class") was a direct attempt to get from pupils exactly what our Test of the Teacher's Ability to Judge Interpupil Preferences was aimed at. The remaining three items in this category were concerned with the teacher's efficacy in promoting the social adjustment and mutual acceptance of all the pupils in her room. Presumably a teacher rated high on these items should have relatively few "isolates," unconscionable "stars," or disruptive cliques.

The "Emotional Appropriateness" Rating-Scale Items

These were Items 3, 6, 9, and 12. In this case, Item 12 ("Knows what worries or bothers you") was designed to tap directly the same kind of perceptiveness as the Problem Prediction Test was intended to measure. The other three items in this group dealt with the kind of emotional supportiveness and acceptance that might be expected of a teacher who was normally motivated to help pupils and who understood their problems well enough to do so.

B. Two Rating-Scale Formats

Experience with rating scales has shown that, unless the raters are carefully trained or other special conditions prevail, there will be fairly high positive correlations among the ratings of all traits or items that have an evaluative connotation. This has been called "halo effect," and it may reflect the perceptual laws delineated by Heider (13). Such a tendency could confidently be expected to operate among our pupils when using the "Our Teacher" rating scale. If so, it would tend to "snow under" any em-

TABLE 7

MEAN UNFORCED RATINGS BY PUPILS ON "OUR TEACHER"

($N = 103$ teachers, 2,885 pupils)

Item No. and Paraphrase	Mean	*SD*	Reliability (Horst)	Intercorrelations Among Items (2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
1. Explains school work clearly	2.41	.28	.76	.25	.66	.65	.65	.42	.60	.58	.60	.33	.56	.50
2. Knows which pupils you like	1.73	.32	.66		.31	.06	.26	.36	.34	.11	.13	.39	.19	.22
3. Makes you feel she likes you	2.07	.35	.71			.67	.66	.54	.65	.61	.66	.29	.58	.63
4. Makes you want to learn things	2.35	.30	.69				.61	.45	.61	.74	.69	.18	.50	.64
5. Makes sure no pupils get left out	2.16	.33	.67					.63	.57	.61	.61	.38	.56	.60
6. Makes sure not to hurt feelings	1.82	.34	.60						.50	.44	.45	.34	.60	.50
7. Catches on to what mixes you up	2.02	.27	.59							.50	.57	.41	.58	.71
8. Helps pupils show what good at	1.94	.34	.70								.69	.17	.49	.62
9. Knows when you are trying	2.22	.28	.61									.25	.53	.62
10. Makes work just hard enough	2.08	.25	.43										.27	.24
11. Sees that no pupils look foolish	1.86	.35	.66											.53
12. Knows what worries you	1.49	.33	.66											
Total	24.17*	2.75*	.79											

* $N = 101$.

pirical support for the logical and psychological distinctions we had attempted to build into the scale. High positive correlations among teachers' mean scores on the 12 items would make it impossible to show that, at least as described by pupils, there are empirically as well as logically distinguishable aspects of teacher effectiveness.

Nonetheless, the over-all "halo" that pupils might have about their teachers could be psychologically meaningful and useful in this research. Such a total mean rating by pupils on 12 positively correlated items would represent the over-all favorability of their attitudes toward the teacher. This total score could serve as a criterion without any of our distinctions among kinds of teacher effectiveness.

To obtain both "halo-affected" and "halo-free" criterion ratings of teachers by their pupils, we collected the ratings in two formats: unforced and forced. We also had, from a previous study of most of the same teachers, scores for an unforced rating scale developed by Leeds.

C. THE UNFORCED RATING SCALE

The unforced format was the "Our Teacher" scale. It is unforced in the sense that the pupil does not have to make distinctions among the items, and halo effect is allowed to operate freely. So a pupil can mark a teacher favorably on all items or unfavorably on all items. We expect the items therefore to be highly positively correlated. Table 7 shows that this expectation was borne out; the median r is .53, with the range extending from .06 to .74.

The internal consistency of the total score on the unforced rating scale was estimated with Cronbach's coefficient *alpha* (4). The obtained value of .92 indicated quite clearly that this rating scale

TABLE 8

FACTOR MATRIX OF 12 ITEMS IN THE UNFORCED RATING SCALE: "OUR TEACHER"

Item No. and Paraphrase	Factor Loadings			Sum of Squared Loadings
	I	II	III	
1. Explains school work clearly	.786	−.045	.176	.6508
2. Knows which pupils you like	.350	.751	−.296	.7741
3. Makes you feel she likes you	.822	−.032	−.041	.6784
4. Makes you want to learn things	.794	−.322	.102	.7445
5. Makes sure no pupils get left out	.807	.019	.006	.6516
6. Makes sure not to hurt feelings	.702	.226	−.358	.6720
7. Catches on to what mixes you up	.803	.106	.038	.6575
8. Helps pupils show what good at	.765	−.300	.047	.6774
9. Knows when you are trying	.780	−.211	.078	.6590
10. Makes work just hard enough	.438	.645	.556	.9170
11. Sees that no pupils look foolish	.733	−.001	−.194	.5749
12. Knows what worries you	.790	−.117	−.081	.6444
Percentage of total estimated common variance accounted for	63.9%	13.0%	6.2%	

involved only one factor of any importance.

Table 7 also shows the means, standard deviations, and reliabilities (15) of the means of pupil ratings of their teachers. The reliabilities indicate the degree to which these means could be expected to correlate with means obtained from hypothetical equivalent samples of pupils. These coefficients range from .43 to .76, with a median of .66. The Horst coefficient for the total score (sum of the 12 means) is .79. Clearly, the "Our Teacher" rating scale elicited descriptions of teachers which differed reliably from one classroom to the next. There is substantial nonerror variance among teachers in the mean rating they receive from their pupils.

Factor Analysis of the Unforced Rating Scale

To check whether, despite halo effect, the items on the "Our Teacher" scale were intercorrelated according to a meaningful pattern, and also to determine the contribution of each item to the general factor, a principal components factor analysis of the intercorrelations in Table 7 was performed. Communalities were used in the diagonals. The sums of squares of the factor loading of the first factors were taken as the communalities in the second, final factoring.

The resulting factor matrix is shown in Table 8. The first factor, a general factor, accounts for 63.9 per cent of the total common variance. We interpret it as the halo factor, or general favorability of the pupils toward their teachers.

The second factor, although it contained 13 per cent of the total common variance, bore no resemblance to any of our postulated dimensions of teacher effectiveness. Its loadings were concentrated in Items 2 and 10, a fact which will assume additional significance when we discuss the factor analysis of the forced rating scale.

Since none of the other factors contained more than 6 per cent of the total common variance, we did not attempt to interpret them.

D. THE FORCED-CHOICE RATING SCALE

The forced format was intended to minimize halo effect, i.e, to secure scores for the teachers on the twelve items that would be relatively unaffected by the pupils' over-all favorability toward their

TABLE 9

MEAN FORCED-CHOICE RATINGS BY PUPILS ON "WHICH IS MORE TRUE OF YOUR TEACHER?"
(N=103 teachers, 2,885 pupils)

Item No. and Paraphrase	Mean	*SD*	Reliability (Corrected* Split-Half)	Intercorrelations among Items (2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
1. Explains school work clearly	7.28	.75	.53	−.39	.20	.46	.31	.12	.21	−.00	.15	.02	−.27	−.29
2. Knows which pupils you like	4.04	.95	.61		−.28	−.19	−.11	−.19	.19	.26	−.03	.44	.17	.50
3. Makes you feel she likes you	5.72	.81	.52			.21	.23	.57	−.31	.17	.22	−.26	.14	−.26
4. Makes you want to learn things	6.21	.70	.30				.09	−.04	.10	.23	.28	.11	−.16	−.24
5. Makes sure no pupils get left out	5.64	.71	.60					.13	.08	−.13	.20	.10	−.04	−.06
6. Makes sure not to hurt feelings	5.06	.62	.45						−.15	.14	.00	−.36	.44	.07
7. Catches on to what mixes you up	5.40	.56	.54							.01	−.03	.30	.07	.31
8. Helps pupils show what good at	5.10	.56	.61								−.02	−.09	.28	.20
9. Knows when you are trying	6.28	.59	.43									.27	−.14	−.04
10. Makes work just hard enough	6.10	.72	.60										−.20	.24
11. Sees that no pupils look foolish	4.55	.74	.54											.36
12. Knows what worries you	4.61	.66	.40											

* Corrected to mean class size, 28 pupils.

teacher. Such scores, when intercorrelated, might more readily reveal a patterning of our 12 items that would empirically corroborate or refine our a priori dimensions (cognitive, social, and emotional) of teacher effectiveness. Further, any group factors that might emerge could be used to obtain factor scores that were empirically, and perhaps educationally, more meaningful than a priori scores.

The forced-choice scale was made by pairing the 12 items in all 66 possible pairs and giving the pupil the following directions:

WHICH IS MORE TRUE OF YOUR TEACHER?

You have answered some questions about how often your teacher does some things. Now you are to choose which of two things is more true about your teacher. On the next few pages are the same questions as those you have just answered. They are in pairs—two at a time. *From each pair you should choose the one thing which is more true about your teacher. Then put a √ mark in the box nearest to that one.*

The teacher's score on any item was the number out of 11 possible choices between pairs in which the pupil chose that item as the more descriptive of his teacher. If, for example, the pupils almost always chose the item "makes sure that no pupil gets left out of things" in preference to the other 11 items with which it was paired, this choice was considered to be evidence that such behavior was highly characteristic of the teacher as seen by these pupils.

The means and standard deviations of the scores obtained by the 103 teachers on each of the 12 forced-choice items are shown in Table 9, along with the corrected split-half reliability estimates of the items. The latter were obtained by correlating (a) the mean number of choices of an item as made by ten of each teacher's pupils with (b) the mean number of choices of that item made by

WHICH IS MORE TRUE OF YOUR TEACHER?

	A more true	B more true	
1A. Makes sure that no pupils get left out of things	*		1B. Makes sure not to hurt your feelings or make you feel afraid
2A. Explains school work so that you can understand it			2B. Knows whom you like best in this class

another ten of her pupils, matched with the first ten as closely as possible according to sex. The resulting r was corrected by the Spearman-Brown formula to estimate the reliability that would characterize teachers' mean item scores based on the ratings of 28 pupils (approximately the mean class size). The reliabilities, while not high, indicated that these scores had possibilities as correlates of other variables.

Obtaining the correlations among forced-choice items presented a special problem. If Item i was to be correlated with Item j, the forced-choice pair formed by Items i and j needed to be disregarded when we obtained the scores for Item i and Item j. Otherwise, an artifactual tendency toward negative correlation would be imposed on r_{ij}, because anyone who chooses Item i in this pair cannot by definition also choose Item j. This meant, of course, getting a different score for each teacher on Item i each time Item i entered into a different correlation. Fortunately, it was possible to handle this heavy scoring and computational burden by means of the Illinois Automatic Computer (Illiac).[4]

The interitem r's in Table 9 are therefore free of any artifactual negative influence. These r's range from $-.39$ to $.57$, the median being .08.

Factor Analysis of the Forced-Choice Rating Scale

The 12×12 correlation matrix formed as described above was factored by the principal components method. Estimates of communality were obtained in two steps. First, the inverse of the correlation matrix was computed to obtain a lower bound estimate of communality from Guttman's equation (11):

$$h^2_i \geq \frac{r^{ii} - 1}{r^{ii}}$$

where r^{ii} stands for the diagonal elements of the inverse of the correlation matrix.

A preliminary factoring with unities in the diagonals indicated that the first five factors would take up most of the variance. The second step in estimating the communalities, therefore, was to get three iterative factorings of the matrix, each time inserting the sum over the first five factors of squares of factor loadings obtained in the preceding factoring. After the third iteration the communalities obtained were used in the final factoring. The resulting factor matrix is shown in Table 10.

Rotation of the factor matrix was carried out by the "quartimax" method (20) to yield the rotated solution shown in Table 11. In this solution there are

[4] The program developed for this work by George C. Stone is on file in the program library of the University of Illinois Computer Group under the title "Analysis and intercorrelation of paired comparisons."

TABLE 10

UNROTATED FACTOR MATRIX OF 12 ITEMS IN THE FORCED RATING SCALE: "WHICH IS MORE TRUE OF YOUR TEACHER?"

Item No. of Paraphrase	Factor Loadings I	II	III	IV	V	Sum of Squared Loadings
1. Explains school work clearly	.545	−.394	.401	−.014	.328	.711
2. Knows which pupils you like	−.691	.112	.213	.101	−.154	.575
3. Makes you feel she likes you	.636	.391	.250	−.051	−.265	.687
4. Makes you want to learn things	.361	−.308	.451	.324	−.052	.536
5. Makes sure no pupils get left out	.233	−.111	.256	−.300	−.001	.226
6. Makes sure not to hurt feelings	.421	.704	.222	−.239	.150	.786
7. Catches on to what mixes you up	−.310	−.243	.362	−.075	.440	.479
8. Helps pupils show what good at	−.067	.360	.376	.594	.008	.604
9. Knows when you are trying	.161	−.197	.339	−.103	−.355	.322
10. Makes work just hard enough	−.496	−.347	.498	−.249	−.270	.739
11. Sees that no pupils look foolish	−.240	.585	.290	−.140	.043	.518
12. Knows what worries you	−.590	.283	.249	−.066	.154	.518
Percentage of total estimated common variance accounted for	34.3%	25.0%	20.3%	10.7%	9.6%	

three large factors, each accounting for about 25 per cent of the total estimated common variance, and two smaller factors accounting for 14 per cent and 10 per cent.

Factor I of the rotated matrix, marked by high loadings for Items 3, 6, and 11, seems to resemble closely our conception of the dimension of *effectiveness in promoting emotional adjustment*. The pupils perceive a teacher who stands high in this factor as making them feel she likes them, making sure not to hurt their feelings or make them feel afraid, and seeing that pupils do not look foolish to the rest of the class. The last item was intended to fall on the dimension of effectiveness in promoting social adjustment. But after the fact, it seems to fit well with the others reflecting concern with the pupils' personal, emotional well-being.

Factor II of the rotated matrix cuts straight across our intended dimensions. It seems to be best interpreted as a rating of *knowledge about the pupils*. It is defined by high loadings in Items 2, 7, 10, and 12. Item 2 is concerned with knowledge in the area of the social structure of the class—the teacher knows which pupils like which others in the class. Items 7 and 10 relate to knowledge about the pupils' difficulties in the

TABLE 11

ROTATED FACTOR MATRIX OF 12 ITEMS IN THE FORCED RATING SCALE: "WHICH IS MORE TRUE OF YOUR TEACHER?"

Item No. of Paraphrase	Factor Loadings I	II	III	IV	V	Sum of Squared Loadings
1. Explains school work clearly	.088	−.002	*.780*	−.079	.314	.721
2. Knows which pupils you like	−.162	*.530*	−.394	.301	−.130	.570
3. Makes you feel she likes you	*.633*	−.266	.332	.085	−.323	.603
4. Makes you want to learn things	−.057	.025	*.667*	.202	−.047	.536
5. Makes sure no pupils get left out	.205	.147	.319	−.237	−.025	.222
6. Makes sure not to hurt feelings	*.881*	−.137	−.006	.036	.075	.802
7. Catches on to what mixes you up	−.117	*.503*	.113	.023	.454	.486
8. Helps pupils show what good at	.166	.051	.014	.773	.008	.628
9. Knows when you are trying	.039	.199	.376	−.048	−.363	.317
10. Makes work just hard enough	−.219	*.790*	.096	−.074	−.251	.750
11. Sees that no pupils look foolish	*.508*	.309	−.332	.204	.007	.505
12. Knows what worries you	.108	*.520*	−.414	.204	.155	.519
Percentage of total common variance accounted for	24.0%	24.9%	27.6%	13.9%	9.7%	

cognitive areas, i.e., whether the teacher knows (according to her pupils) what mixes the children up in school work, and whether she makes school work hard enough but not too hard. Finally, in Item 12, this factor involves knowledge about the pupils' emotional adjustment—i.e., whether the teacher knows what worries or bothers her pupils.

Factor III seems plainly to be related to the dimension of *effectiveness in promoting cognitive adjustment*. Very high loadings were obtained for Items 1 (explains school work clearly) and 4 (makes pupils want to learn new things). The factor is also marked by moderately high loadings for several other items, positive for 3, 5, 9, negative for 2, 11, and 12. These loadings seem to complicate the meaning of this factor. It may be that this factor represents the "conventional virtues of a teacher." Its high loadings are on items on which pupils and laymen would expect teachers to have high scores. Its negative loadings are on items which, however significant to educational psychologists, are not currently accepted by pupils as established requirements of the teacher's role.

This interpretation is supported by the following evidence: There is a rho of .93 between the loadings of the items on Factor III and their difficulties in the forced format (i.e., their mean scores as shown in Table 7). Thus the more often an item was chosen in preference to other items, the higher its loading on Factor III. The more often it was chosen, the more it was seen by the pupils as true of their teachers in general. The latter also means, it seems reasonable to say, "the more it was expected and hence observed."

Factor IV has high loadings on Item 8 (helps all pupils show what they are good at), and to a considerably less extent Items 2 (knows which pupils you like best in the class) and 4 (makes you want to learn new things). Factor V is defined by positive loadings on Items 7 (catches on quickly to what mixes you up in school work) and 1 (explains school work so that you can understand it), and negative loadings on Item 9 (knows when you are trying hard) and Item 3 (makes you feel she likes you). Because these two factors accounted for little of the total common variance, they were not used to obtain factor scores.

Teachers' Scores on Three Forced Factors. Since the first three factors seemed to be fairly interpretable, factor scores were computed for all 103 teachers on each of these three, according to the method described by Holzinger and Harman (14, pp. 267 ff.).

TABLE 12

FACTOR SCORES FOR RATING ON FORCED RATING SCALE: "WHICH IS MORE TRUE OF YOUR TEACHER?"
(N=103)

Factor	Mean	SD	Intercorrelations	
			II	III
I	.014	.871	−.37	−.24
II	.011	.817		−.34
III	−.003	.871		

In theory, factor scores calculated in this manner should be uncorrelated standard scores. However, in the present case, because of the special method used in obtaining the correlation coefficients, the scores are actually negatively correlated to some extent, and do not have exactly the means and standard deviations usually expected of standard scores. It will be recalled that the correlation coefficients were calculated by using a different score for each item each time it was paired with another. In obtaining the factor score it was, of course, necessary to use a single score for each item. Thus the bias for which we corrected in the calculation of the correlation coefficients had to be allowed to re-enter in the factor scores. The means, standard deviations, and intercorrelations of the three factor scores are given in Table 12.

E. LEEDS' "MY TEACHER" INVENTORY

A criterion was also available from a study (6) performed in the same school system the previous year. These were the teachers' scores on Leeds' "My Teacher" Inventory. This has 50 items answered Yes, No, or ? by the pupil, and scored 1, −1, and 0 on each item. Eighty of our 103 teachers had such scores. The mean and standard deviation of their scores were 21.76 and 9.38, respectively. The

TABLE 13
CORRELATIONS OF SCORES ON ITEMS

	No. of Teachers	Item											
		1	2	3	4	5	6	7	8	9	10	11	12
1. *r* of forced item with unforced total	103	.25	−.32	.33	−.07	.14	.34	−.04	−.15	−.22	−.30	.11	−.10
2. *r* of unforced item with unforced total	103	.78	.41	.84	.78	.82	.73	.80	.76	.78	.47	.79	.79
3. *r* of unforced item with forced item	103	.30	.37	.51	.18	.32	.36	.15	.09	−.04	.21	.22	−.02
4. *r* of unforced item with Leeds inventory	80	.25	.03	.25	.17	.19	.10	.15	.15	.17	.11	.39	.25
5. *r* of forced item with Leeds inventory	80	.04	.03	.25	−.07	.06	.14	−.15	−.07	.07	.14	.21	.04

Horst reliability of the scores obtained for the 97 teachers originally tested was .91; by estimating from the range of talent in that group to the range of talent in our group, we obtained a reliability estimate of .90. No new reliability coefficient was calculated.

F. Relationships between Forced and Unforced Items

Item Difficulty

How did the mean values of the 12 items on one format correlate with those on the other? This question bears on the degree to which the "difficulty" of an item, or the ease with which teachers could satisfy its requirements in pupils' eyes, was consistent from one format to the other. The "easier" an item, the higher the mean score on that item on both the unforced and forced scales. This would hold, however, only insofar as the pupil's ratings on the scales were nonrandom and consistent. The rank-difference correlation between the two sets of 12 means from Tables 7 and 9 was .94. Thus there was high consistency of item "difficulty" from one format to the other.

Item Favorability

The over-all favorability of the children toward their teacher can be defined as her total score on the unforced rating scale. Then the correlation of the teachers' scores on any item with their total scores on all items is an estimate of the favorability of that item.

The forced items differ in favorability. These differences are revealed in the correlations with the total score on the unforced scale, shown in Table 13. It is noteworthy that Items 2 and 10 are much lower in favorability than the rest. We consider these differences in favorability within the forced rating scale in the next section.

What is the relationship between difficulty and the favorability of the forced items? The rank-difference correlation of .15 indicates that the difficulty of the forced rating items has but little if any relation to their favorability.

We can also compare the favorabilities of the items in the two formats. We do this by correlating (*a*) the correlations of the unforced item scores with the total scores on the unforced rating scale with (*b*) the correlations of the forced item scores with the same total. The rank-difference correlation of .50 indicates that the agreement between the favorabilities of the items in the two formats is only moderate.

Correlations Between Teachers' Scores on Corresponding Forced and Unforced Items

How did the teachers' scores for a given item on one rating scale correlate with their scores for the same item on the other scale? These *r*'s, shown in Row 3 of Table 13, range from −.04 to .51. Although all but two of these *r*'s were positive, they were fairly low, their median being .22. Apparently the meanings of the item scores changed from one format to the other. In the unforced format, halo or over-all favorability largely determined the pupil's judgment on each item. In the forced format, the pupil had to choose between two favorable statements about his teacher.

Why did the corresponding items on the two formats not correlate more highly? Let us assume that the children are still attempting to respond with a favorability (not necessarily a "favorable") rating of their teacher even on the forced scale. Insofar as favorability is the primary basis for rating the teacher, and insofar as item favorability differs from one format to the other, the response to the item on the two formats may be expected to differ. There should then be a negative relationship between (*a*) the difference in favorability of the item from the unforced to the forced format, and (*b*) the correlation between scores for the same item on the two formats. Thus we can rationally rank correlate the difference between Rows 1 and 2 with the value of Row 3 in Table 13. The rho of −.91 between the amount of difference in correlation with the total (unforced) "Our Teacher" score and the interformat item correlations indicates that almost all of the change in meaning of the items as the format changed is related to change in favorability.

A more direct comparison of the favorabilities of the items is possible. As suggested above, we can estimate the favorability which is expressed by an item score from the correlation of that score with the total score on the "Our Teacher" unforced rating scale. Table 13 shows that in general the relevance of the forced items to favorability was lower than was the relevance of the corresponding unforced item. We expected this result, for we tried to eliminate the favorability factor, or halo, in the forced format. But there is no apparent reason why the relative favorabilities of the items should change. If an item is low in favorability in the unforced scale, it should become unfavorable when a discrimination between items is required. Items 2 and 10, which were much lower in relevance to favorability in the unforced scale than were the other items, came out, as expected, most unfavorable in the forced scale. But the over-all agreement between relative favorability of the items in the two formats is not high. The rank-difference correlation between the favorability of an item in the unforced scale and the favorability of the

TABLE 14

INTERCORRELATIONS AMONG THE FIVE COMPOSITE RATING SCORES
($N = 103$, except as noted)

	Score			
	2	3	4	5
1. "Our Teacher" total	.37	−.25	.14	.25*
2. Forced Factor I (Affective)				.14*
3. Forced Factor II (Informative)				−.05*
4. Forced Factor III (Cognitive)				−.04*
5. Leeds' "My Teacher" total				

* $N = 80$.

same item in the forced scale is only .50. This low rho helps us to understand the low correlations obtained between the item scores from one format to the other.

G. Relationships between Item Scores and Scores on Leeds' Inventory

If the analysis of item favorability described above is sound, there should exist relations between the item scores and the Leeds inventory total score similar to those we have already discussed. Table 13 shows the correlations obtained. Of course we should expect some changes, since as we shall discuss later, the relationship between the Leeds inventory total score and our unforced inventory total score was only .25. In general, the correlations of the items with the Leeds score are smaller than those with the "Our Teacher" scale. These item correlations against the Leeds scale would be reduced both by differences between the Leeds and "Our Teacher" scales and by the fact that the scales were used a year apart by different sets of pupils. Rank-difference correlations between the estimates of item favorability, obtained by comparison with the two different total scores, were positive, however. The rho (Row 2 vs. Row 4 of Table 13) for the unforced items (Leeds favorability vs. "Our Teacher" favorability) was .54; the rho (Row 1 vs. Row 5 of Table 13) for the forced items was .40. Thus it appears that the Leeds and "Our Teacher" scales agree somewhat as to the meanings of both unforced and forced items.

The relation between the Leeds favorability of the items in the two formats (Row 4 vs. Row 5 of Table 13) is expressed by the rho of −.02. Also, what is the relation between the amount of difference in Leeds favorability of an item between unforced and forced formats and the correlation between scores for the item obtained in the unforced and forced formats? The rho in this case is −.58, to be compared with the value −.91 obtained earlier. Thus, the whole pattern of relationships between item scores and the total favorability score seems to bear up under the change of favorability score from "Our Teacher" to Leeds, even though the magnitudes of the correlations are in all cases reduced when the previous year's Leeds scores are used.

H. Relationships Among Composite Rating Scores

We have described five composite rating scores: the total score on the "Our Teacher" inventory, three factor scores from the forced rating scale, and the score on the Leeds inventory given in the previous year. Table 14 shows their inter-

TABLE 15

SEX DIFFERENCES IN TEACHERS' RATINGS BY PUPILS

Rating	Male Teachers ($N=19$)		Female Teachers ($N=84$)		Difference in Means
	Mean	*SD*	Mean	*SD*	
Unforced Total	23.72	3.165	24.28	2.604	−.56
Factor I	2.578	1.080	3.097	.785	−.519*
Factor II	2.659	.715	1.853	.762	.806**
Factor III	3.875	1.119	4.029	.801	−.154

* Significant at the .05 level.
** Significant at the .001 level.

correlations. (The negative correlations among the forced factors, having already been presented and discussed, are omitted here.) The correlation of .25 between the scores on the two unforced rating scales is disappointingly low. It is just significant at the .02 level (one-tailed test); corrected for attenuation, the r becomes .29. This correlation may be low because more than a year elapsed between the two ratings, inasmuch as different classes of pupils made them. The foregoing discussion of changes in item-total r's when the total score was changed from that of the "Our Teacher" rating scale to that of the Leeds inventory, suggests, however, that there are also some differences in the meanings of the items in the two scales.

The correlations of the factor scores with the totals on the unforced scales indicate that teachers high in the factor concerned with personal adjustment (Factor I) are positively valued. Those high in the "cognitive" factor (Factor III) are also positively valued when this year's ratings are considered, but to a lesser degree. Teachers high in the factor of "perceived knowingness" (Factor II) are negatively valued. The correlations of the factor scores with the Leeds scale are close to zero.

I. PUPILS' RATINGS OF TEACHERS IN RELATION TO OTHER CHARACTERISTICS OF THE TEACHER

What are the relationships between pupils' ratings and the four characteristics of teachers mentioned at the end of Section II: sex, age, grade level taught, and score on the Minnesota Teacher Attitude Inventory?

As shown in Table 15, female teachers were given slightly higher over-all favorability ratings (on the unforced scale) than were male teachers, but the difference was not significant. Females were also rated higher on Factor I ("Affective merit"), and this difference was significant at the .05 level. Male teachers, on the other hand, were much higher ($p < .01$) in Factor II, the "inquisitiveness" or knowledgeability factor. There was almost no difference in ratings received by male and female teachers on Factor III. Thus, the women teachers seem to be regarded as warmer and kinder by children, while the men are rated as "knowing" more about the pupils.

Both rectilinear and curvilinear relationships between the age of the teachers and their ratings by pupils were essentially zero. (Curvilinear relationships were determined by analysis of variance among mean ratings of teachers in three age groups: 22-32, 33-48, and 49-62.)

Sixth-grade teachers are seen by their

TABLE 16

GRADE DIFFERENCES IN TEACHERS' RATINGS BY PUPILS

Rating	Fourth Grade ($N=34$) Mean	*SD*	Fifth Grade ($N=33$) Mean	*SD*	Sixth Grade ($N=36$) Mean	*SD*	*F*	Differences in Means 4–5	5–6	4–6
Unforced Total	24.32	2.22	24.66	2.64	23.59	3.10	1.34			
Factor I	3.29	.77	3.11	.70	2.63	.97	5.77**	.18	.48*	.66**
Factor II	1.46	.50	2.06	.91	2.46	.66	17.16**	−.60**	−.40*	−1.00***
Factor III	4.24	.67	3.78	.96	3.98	.90	2.44			

* Significant at the .05 level.
** Significant at the .01 level.
*** Significant at the .001 level.

pupils as significantly less supportive (Factor I) than are teachers in Grades 4 or 5 (Table 16). On the other hand, the Factor II ratings received by sixth-grade teachers are significantly higher than those of fourth- or fifth-grade teachers. Fifth-grade teachers were rated lowest on Factor III (Cognitive Merit) but only the difference between fourth- and fifth-grade teachers was significant.

The decrease in Factor I ratings from Grade 4 to Grade 6 might arise (*a*) from changes in the pupils' attitudes as they progress through the intermediate grades; (*b*) from changes in the entire classroom situation—roles, expectations, etc., or (*c*) from selection of more supportive teachers for the lower grades. The last possibility would be somewhat refuted by the fact that mean scores on the Minnesota Teacher Attitude Inventory were progressively higher (although not significantly) for each successive grade level from fourth to sixth. If MTAI scores reflected affective merit and supportiveness, we would have to reject the hypothesis of selective placement of supportive teachers in the lower grades. But it is shown later (p. 27), however, that teachers' scores on the MTAI were uncorrelated with their pupils' rating of their "affective merit." This means that the increase from Grade 4 to Grade 6 in MTAI mean scores may be irrelevant in explaining the differences among grades in Factor I ratings.

Thus, women teachers and teachers of the fourth grade are rated high on the affective factor, and male teachers and sixth-grade teachers are high on the inquisitiveness or knowingness factor. Since almost all of the male teachers were in

TABLE 17

GRADE DIFFERENCES IN FEMALE TEACHERS' RATINGS BY PUPILS

Rating	Fourth Grade ($N=34$) Mean	*SD*	Fifth Grade ($N=29$) Mean	*SD*	Sixth Grade ($N=21$) Mean	*SD*	*F*	Differences in Means 4–5	5–6	4–6
Unforced Total	24.32	2.22	24.66	2.71	23.66	2.90	.87			
Factor I	3.29	.77	3.15	.63	2.73	.87	3.57*	.14	.42	.56*
Factor II	1.46	.50	1.99	.89	2.30	.60	10.42**	−.52**	−.31	−.84***
Factor III	4.24	.67	3.83	.84	3.95	.86	2.17			

* Significant at the .05 level.
** Significant at the .01 level.
*** Significant at the .001 level.

the sixth grade, and none was in the fourth, the differences in ratings associated with grade might well have been due to the sex factor. We have recalculated the ratings by grades for female teachers only. The results are shown in Table 17. The same trends from Grade 4 to Grade 6 appear, although the significance of the differences is reduced in some cases.

Since Factor II, the knowingness factor, was evaluated somewhat negatively by pupils, it appears that children may be less satisfied by the instruction they receive at the upper elementary-grade levels than they are at lower grade levels. Differences between male and female teachers do not appear to be important, when grade level is taken into account.

IV. RELATIONSHIPS BETWEEN TEACHERS' PERCEPTIONS AND BEHAVIORS

This section presents the relationships between the measures of teachers' accuracy and the measures of their behavior as described by pupils.

Most of the correlations are essentially zero. This is true of each of the tests in relation to (*a*) the total scores on the "Our Teacher" rating scale, and (*b*) the scores on the logically most relevant items in both the unforced and forced rating scales.

For each test we have also considered all of its *r*'s of .20 or above with other variables. Some of these *r*'s are based on only 80 cases; these 80 teachers were all those on whom data were available not only on our measures but also on those used in a research project (6) carried out the year before (the Minnesota Teacher Attitude Inventory, the Leeds "My Teacher" rating scale, and an Inventory of Pupils' Cognitive-Affective Values). None of the correlations between MTAI scores of a year before and pupils' ratings of the same teachers a year later was significantly different from zero.

The correlation between teachers' accuracy in judging interpupil preferences and ratings on Unforced Item 2 (Knows whom you like best in this class) is .28. An *r* of this magnitude based on 103 cases is ordinarily considered significantly greater than zero. Similarly, the highest *r* of this regression-corrected ratio score with forced-choice rating scale items was that with Forced Item 2 ($r = .20$). Special attention is called to these correlations not only because they were significant but also because they seemed meaningful. The rating-scale item involved here is the one that was explicitly written to secure descriptions by pupils of the very same characteristic of teachers as was measured by the test. The fact that this item and the test correlate not only significantly but more highly than either variable does with any other seems to be evidence for the validity of both variables. It should also be noted, however, that the correlational score on this test had lower *r*'s (.10 and .06) with this item of the unforced and forced rating scales.

It seemed possible that the relationship might be explained in part by the teacher's use of an easily recognizable method for obtaining knowledge of pupil preferences. That is, her administration of a sociometric test in the class might both increase her knowledge of pupils' choices and cause the children to believe that she knew whom they liked best. Accordingly, we separated into two groups the 93 teachers who had answered

TABLE 18

DIFFERENCES BETWEEN TEACHERS WHO HAD AND HAD NOT OBTAINED SOCIOMETRIC INFORMATION FROM THEIR PUPILS

Variable	Teachers who had obtained sociometric information ($N=24$)			Teachers who had not obtained sociometric information ($N=69$)				
	Mean	*SD*	r with rating on unforced Item 2	Mean	*SD*	r with rating on unforced Item 2	Difference in means	t
Accuracy in predicting interpupil preferences								
Regression-corrected ratio score	1.303	.195	.315	1.212	.204	.185	.091	1.90*
Correlational accuracy score	.534	.167	.085	.456	.194	.114	.078	1.70*
Mean rating on unforced Item 2	1.814	.361	—	1.684	.303	—	.130	1.69*

* Significant at .05 level (one-tail test).

our question, "Have you ever asked your pupils in this class to write down their preferences among their classmates? Answer 'yes' only if you have had them do it this year." Twenty-four teachers replied "yes" and 69 "no." Table 18 compares the two groups as to both kinds of scores for ability to predict pupils' sociometric choice, as to mean ratings on Unforced Item 2, and as to correlations between accuracy and mean rating. Teachers who had given sociometric tests were significantly more accurate in predicting pupil choices than those who had not given such a test. They were also rated higher on the question "Knows whom you like best in this class." As would be expected, removing this source of covariance reduced the correlation between accuracy and rating, bringing it in all cases below the .05 significance level. Nevertheless, the correlations remained positive in all cases, and the joint probability[8] of the two r's for the regression-corrected ratio score is less than .05. It is therefore probable that there are other cues than the giving of sociometric tests available to the children which enable them to evaluate their teacher's knowledge of their sociometric preferences.

The correlation between teachers' accuracy in judging interpupil preferences and their pupils' mean socioeconomic status is .27. The latter variable was measured with a five-item inventory that asked whether the pupil's family had a vacuum cleaner, an electric or gas refrigerator, a bathtub or shower with running water, and a telephone, and whether the pupil had ever had paid lessons in dancing, art, music, etc., outside of school. The r of .27 indicates that teachers were more accurate if their pupils came from homes that had higher socioeconomic status. This relationship has several possible meanings: (*a*) Higher status pupils may be more active in interpupil relationships, so that their interpupil preferences are easier to discern. (*b*) Classrooms containing higher status pupils may have greater social cleavage,

[8] Joint probability was estimated by the chi-square method described by Jones and Fiske (16).

making their sociometric structure easier to perceive. However, we found a slight negative r (−.18) between mean socioeconomic status and the class's standard deviation of sociometric choices. (*c*) Higher status pupils have preferences more similar to those of their teachers and hence more easily predicted by the teachers. These and other possible interpretations can be tested empirically in part with data already collected.

The correlation between teachers' accuracy in predicting pupils' problems and ratings on Unforced Item 2 (Know whom you like best in this class) is .29. This r seems meaningful because the Problem Prediction Test is positively related (.33) to the test of accuracy in judging interpupil preferences.

The correlation between teachers' accuracy in predicting pupils' problems and their pupils' mean socioeconomic status is .27. This coefficient suggests that teachers, whose own backgrounds are generally assumed to be "middle class," are better at understanding the problems of pupils of similar background. If this is so, we should also expect a relationship between teachers' "liking of pupils" and their accuracy in predicting the pupils' problems. "Liking" is assumed to be indicated by the teacher's choice of the pupil as one she finds "easiest to work with" and "dislike" by choice as "most difficult to work with." The difference between teachers' mean scores in predicting the problems of these two groups of pupils was in the predicted direction with a critical ratio of 1.77, significant at about the .04 level by a one-tail test.

The correlation between teacher's Factor II rating ("knowledgeability") and the mean socioeconomic status of the class is .37. Higher status children thus tended to rate their teachers more highly on knowing whom they like best, what their problems are, and what mixes them up in school work.

A. Curvilinear Relationships

A very low or zero Pearson r between two variables does not, of course, preclude the possibility of a significant curvilinear relationship between them. Although we had not expected curvilinear relationships, the possibility of their occurrence seemed worth investigating. Accordingly, we divided the group of teachers into fourths, using scores on each of the various tests and composite ratings in turn, as the basis for the breakdowns. The means of the four groups' scores on the dependent variables were then inspected. Wherever the mean differences seemed to justify it, coefficient epsilon was computed to estimate the closeness and significance of the relationship, curvilinear or otherwise, that might exist between the two variables. No significant epsilon values were found.

B. Relationships Between Patterns of Test Scores and Ratings

Despite independence between single variables, patterns of scores on the variables might be significantly associated. To test this possibility the scores on each of the three tests of understanding, here referred to as cognitive (C), affective (A) and sociometric (S), were standardized and used to group the teachers according to the pattern of their scores. For example, a teacher whose standard scores on Tests C, A, and S are +1.5, 0, and −1.5, respectively, would be considered to have a test-score pattern of CAS. If another teacher's test scores were −1.5, 0, and 1.5, respectively, her pattern would be called SAC. The teacher's pat-

TABLE 19
CONTINGENCY TABLE—TEST PATTERN VS. RATING PATTERN

Test Pattern	Rating Pattern						Sum
	ACS	ASC	CAS	CSA	SAC	SCA	
ACS	2	1	3	2.5	1	4	13.5
ASC	0	0	3	3	1	8	15
CAS	7	1	5	2.5	1	8	24.5
CSA	5	3	3.5	3	3	2	19.5
SAC	2	5	3	4	3	1	18
SCA	3	2	0.5	0	4	3	12.5
Sum	19	12	18	15	13	26	103

tern is thus set by the rank order of her standard scores on the three tests.

Similarly, the factor scores for the ratings received by the teacher from her pupils were used to group the teachers according to patterns. A 6×6 contingency table shown in Table 19 was made to show how the teachers' patterns of test scores and ratings were distributed. In general, the distribution of patterns of test scores alone did not depart significantly from that to be expected by chance. This was also true of the distribution of patterns of ratings alone. Finally, the joint distribution, or contingency table, did not reveal significant differences, in pattern frequency, from those to be expected by chance.

C. Second-Order Relationships

By a second-order relationship we mean one in which a relationship, e.g., a correlation, between two variables is a function of a third. Thus, in one study (6) it was found that the correlation between the Minnesota Teacher Attitude Inventory and pupils' ratings of teachers on the Leeds "My Teacher" inventory was a function of a third variable, pupils' cognitive-affective values as to the kind of teacher merit they considered most important.

To investigate whether second-order relationships existed in the data of the present study, we examined the correlations between our tests and pupils' ratings of teachers as a function of four different second-order determiners: sex of teacher, age of teacher, teacher's score on the Minnesota Teacher Attitude Inventory, and grade level. The distributions of teachers' ages and the MTAI scores were divided at the quartiles. Correlations were calculated separately for the highest, the second and third combined, and the lowest fourths. Separate correlations also were computed for teachers of each sex and of each grade level (fourth, fifth, and sixth) after mixed classes were included with the higher of the two grades involved.

Sex and age

None of the differences between correlation coefficients for male and female teachers was significant. Age groups differed only with respect to the correlation between accuracy in predicting sociometric preferences and the affective factor. For the 26 youngest teachers (median age 26) this correlation was $-.129$, and for the 26 oldest teachers (median age 53) it was $+.395$. The difference between these r's is significant at the .05 level.

MTAI Score

For the 20 teachers who were lowest and the 40 who were in the middle range on MTAI, there were positive r's (.23 and .20) between accuracy in predicting sociometric preferences and the unforced rating of the pupils. However, for the 21 teachers highest on MTAI there was a substantial negative r (−.42). Both positive correlations differed from the negative one at the .05 level of significance. Differences between MTAI groups were in the same direction, and significant at the same level of confidence, when the r's between WQHT[6] and the unforced total rating were compared and when the r's between sociometric accuracy and Factor I were compared. In each case, however, the signficant difference ($p < .05$) between correlations was observed only between the middle group and the high group on MTAI. Interpreted literally, these r's suggest that understanding of pupils is valuable in helping teachers get favorable ratings from pupils when the teachers have average scores in MTAI, but understanding may be detrimental to teachers with high MTAI scores.

Grade

Grade level as a second-order variable yielded significant differences in correlation as follows. There was a consistent decline in the r between the teacher's WQHT[6] score and Factor III rating from the fourth (r, .38) through the fifth (r, −.04) to the sixth grade (r, −.29). A similar decline was observed in the correlation between accuracy in predicting pupils' problems and Factor I rating; the r's in the fourth, fifth, and sixth grades were .17, −.10, and −.35, respectively. The r's between problem prediction accuracy and Factor II ratings had an opposite trend from Grade 4 to Grade 6: −.30, −.05, and .22, respectively. Finally, the r's between problem prediction accuracy and Factor III ratings showed a curvilinear trend from Grade 4 to Grade 6: .36, −.27, and .25. The general picture here is one in which the teacher's accuracy in predicting pupils' problems becomes associated with less favorable ratings by pupils as we go from lowest to highest in the three grades studied.

V. IMPLICATIONS FOR FURTHER RESEARCH

The results reported in Section IV provide little support for the proposition that teachers should understand their pupils. With one exception, the correlations between measures of teachers' understanding and pupils' ratings of their teachers were insignificant. No curvilinear relationships have thus far been found, nor significant relationships between patterns of test scores made by the teachers and patterns of factor scores based on pupils' ratings. Various second-order relationships have emerged, but these are as yet only suggestive.

From this paucity of positive findings, it is, of course, too hasty a conclusion that teachers' understanding of their pupils is unrelated to any valued phenomena. There are at least three other possible conclusions compatible with these results: (*a*) The present tests, however plausible, are nonetheless irrelevant to teachers' understanding of their pupils. (*b*) The pupils' ratings are in-

[6] "Which Question is Harder?" test.

valid as measures of their attitudes toward their teachers. (*c*) The present tests measure understandings related to kinds of effectiveness other than those obtained with pupils' ratings.

Have we made progress toward delimiting the terms of the original proposition and specifying their dimensions? This question can best be answered by evaluating each of our instruments in turn, and suggesting modifications which may be fruitful. This procedure may also point to new directions for research on social perception in the classroom. The evaluation of the measures of understanding depends in part on their relations to the pupils' ratings. We therefore consider the rating scales first, in order to establish a basis for judging our tests of understanding.

A. The "Our Teacher" Inventory

What aspect of the children's perceptions is measured by the "Our Teacher" inventory? This question is difficult to answer from our data. We have assumed throughout this report that the total score was a measure of favorability of attitude, or general approval of the teacher. Since the children were not asked specifically how well they liked their teacher, this assumption could not be tested. The only ground for doubt is the relatively low correlation (.25) with the Leeds inventory, another internally consistent rating scale.

If the two scales actually measure different aspects of the children's perceptions then it is not possible that both are measuring pure favorability. A study is needed in which the "Our Teacher" inventory, Leeds' "My Teacher" inventory, and a third scale or inventory, designed to measure pure favorability toward the teacher, are all given to the same groups of pupils. The same instruments should be given to pupils of the same teachers in the next year. Such a study should provide evidence for the evaluation of two assumptions: (*a*) that the general factor or total score in our unforced rating scale is a favorability factor, and (*b*) that favorability of pupils toward their teacher is largely due to characteristics of the teacher and relatively independent of the particular children involved.

B. The "Which Is More True of Your Teacher" Inventory

Our evaluation of the forced-choice rating scale must be more cautious than was our judgment of the unforced scale. Item reliabilities, while lower than those for the same items in the unforced format, are high enough to indicate a considerable proportion of non-error variance. Our effort to eliminate the favorability factor by the forced-choice format was not entirely successful, as indicated by the significant correlations of Factors I and II with the total score on the unforced scale. But these correlations seem low enough, pending an estimation of the reliability of the factor scores, to indicate that something other than favorability is being measured in the factors.

While we have interpreted each of the three major factors in terms of the items which are most highly loaded, we feel no assurance that these interpretations actually correspond to the attitude or perception being measured.

The *change* in favorability of items, as the format changes, certainly raises doubts. Why should a perceived teacher characteristic be highly correlated with the over-all favorable attitude toward the teacher when considered alone, but become uncorrelated or negatively cor-

related with this over-all attitude when considered in conjunction with some other teacher characteristic? We suspect that, in choosing between characteristics, many of the children were no longer judging whether their teacher actually displayed the behavior in question. The forced-choice technique may not be workable with children at this early age. Our suspicions are supported by the fact that scores on the items do not correlate highly from one format to the other.

If the item scores cannot be accepted at face value, what of our interpretations of the factors? We find that the loadings of the items on Factors I and II correlate (rho) .79 and −.79, respectively, with the favorability of those items. Further, the loadings of the items on Factor III correlate (rho) .93 with the frequency with which they were chosen in preference to other items (i.e., item "difficulty"). What are the two orthogonal factors which have opposite correlations with favorability and whose item loadings correlate −.77 with each other? What is the factor whose item loadings have almost perfect correlation with the item "difficulties"? Further investigation—by comparing scores on these factors with the ratings of outside observers, and by having the children and teachers respond to the items according to various sets of instructions (e.g., "How important is this in a teacher?", "How much do you like this in a teacher?")—might be rewarding.

In short, the present forced rating scale, although interesting and suggestive, yields scores about whose meaning serious question can be raised. It can be sharpened in two ways: by collecting data to clarify the meaning of the present form, and by revising items to equalize their favorability and "difficulty."

C. The "Which Question Is Harder" Test

It is possible that this test was designed in a way which cancelled out one important aspect of cognitive understanding, and failed to get at another. First, our procedure prevented any measure of knowledge of the *absolute* difficulty of intellectual tasks. Yet such knowledge may be more relevant to effective teaching than knowledge of *relative* difficulty.

Second, our decision to score the test against national norms ruled out any measure of the teacher's knowledge of her own pupils' abilities. Experience with the order of presentation of subject matter in texts and tests could lead to knowledge of the relative difficulty of the material on a nationwide basis. But knowledge of the relative difficulty of the items for her own class might prove to be more valid as a correlate of the teacher's effectiveness in the eyes of her pupils.

Giving subject-matter tests to obtain a scoring key, however costly of testing time, would make possible the latter kind of test. Also, either a nationally or locally keyed test could be administered in an unforced form. That is, we could ask teachers to estimate the actual proportion of pupils who would pass each item. We could compare the difficulty, reliability, and intercorrelation of two forms of such a test: one with "anchor" items of whose difficulty the judges would be informed (19), the other without such items.

D. The Test of the Teacher's Ability to Judge Interpupil Preferences

Although teachers' scores on this test did not correlate significantly with pu-

pils' over-all evaluation of the teacher, it had more significant and meaningful relationships with other variables than either of the other two accuracy tests. Of the two different scores obtained with this test, the correlations are without exception higher for the regression-corrected ratio score based on prediction of specific choices; hence, we regard this as the more promising measure.

How shall we account for the absence of correlation between the teacher's accuracy in judging interpupil preferences and the pupils' over-all favorability toward her? Is it that the teacher's understanding of sociometric structure is not related to her effectiveness in promoting social adjustment? Or is it that the more effective teacher has reduced the "spread" of sociometric status in her class by reducing isolation and stardom, and so has made the prediction task more difficult for herself? Or is the children's favorability toward the teacher unrelated to her effectiveness in promoting social adjustment? Decision among these alternatives must wait on the development of an "objective" measure of teachers' effectiveness in promoting social adjustment.

E. The Problem Prediction Test

The accuracy score on this test had a corrected split-half reliability of only .26. It correlated .33, however, with teachers' accuracy in predicting interpupil preferences. Hence, insofar as this test measured anything, it seemed to be measuring some sort of understanding of pupils by teachers. Its other significant correlations are with variables which also correlated significantly with accuracy in predicting interpupil preferences.

The Problem Prediction accuracy score may be no more than a less reliable test of the same ability that is measured by the Test of the Teacher's Ability to Judge Interpupil Preferences. The fact that two such different tests as these have such similar relations to other variables suggests that both reflect a more basic variable.

The low reliability of the test may well have been due in large part to the difficulty of the task set for the teachers. If it had been easier, i.e., if less extrapolation from teachers' knowledge had been necessary, higher reliability probably would have resulted. One way to make the task easier may be to make it a free-choice rather than a forced-choice task. This would permit teachers to judge whether a pupil had a given problem as against whether the problem was greater or less than some other problem.

F. Conclusions

The final outcome of the analyses reported in this paper may be stated as a decision with regard to the future of each of the five instruments.

1. *The "Which Question Is Harder?" Test* proved moderately reliable and positively correlated with teaching experience. It needs further exploration of its relationship to other measures of cognitive understanding of pupils and to teacher behavior; other approaches to the measurement of accuracy in understanding cognitive aspects of pupils need to be explored.

2. *The Test of the Teacher's Ability to Judge Interpupil Preferences* proved moderately reliable and had meaningful correlations with other variables; it should be validated against additional criteria.

3. *The Problem Prediction Test* holds little promise in its accuracy score. Inso-

far as it measures anything it seems to measure whatever is tapped by the Test of Ability to Judge Interpupil Preferences, but with less reliability.

4. *The "Our Teacher"* rating scale has properties which are now better known than most comparable devices for use in Grades 4-6. Already usable for appraising pupils' general favorability toward their teacher, it can be made even more so by modifications in the light of the factor analysis.

5. *The "Which Is More True of Your Teacher?"* forced rating scale has moderate reliability and fairly meaningful factorial structure. It is not yet well enough understood in its present form for use as a criterion instrument. Our results indicate that further research can provide a tool for obtaining differential, "halo-free," ratings of teacher behavior by pupils.

VI. Summary

To analyze and test the general proposition that teachers should understand their pupils, three tests of teachers' understanding of pupils were devised and administered to 103 teachers of fourth, fifth, and sixth grades. These tests were correlated with pupils' descriptions of teacher behavior on both a forced-choice and an unforced-choice rating scale. On a priori grounds, we selected three areas from the domain of educational concerns for testing these relationships: (*a*) cognitive, (*b*) social, and (*c*) personal problems.

To measure their understanding of cognitive aspects of pupils, we presented the teachers with 60 pairs of achievement-test items. Each pair differed in the percentage of pupils of a national sample who answered correctly. The teacher was asked to indicate, for each pair, which item was more difficult for pupils in the fourth to sixth grades. The teacher's score was the number correct out of 60.

To measure the teacher's awareness of the sociometric structure in her class, she was asked to "predict" which two children each of her pupils would prefer to have in the same section if the class were split into two sections. The number of correct predictions, divided by the number of pupils in the class, yielded the mean number of "hits" per pupil. This score was corrected by its regression on class size to avoid giving an advantage to teachers with small classes. In addition, the r between predicted and actual choice status of her pupils was computed for each teacher as a correlational accuracy score.

To measure the teacher's sensitivity to her pupils' personal problems, we made up 12 sets of three-problem check-list items equated for prevalence in a national sample. The pupils were asked to rank the three items in each set according to how much each "worried or bothered" them. Each teacher was asked to predict the problem-ranking responses of the two boys and two girls who were "easiest to work with" and two boys and two girls who were "most difficult to work with." The teacher's accuracy score was the mean (over the eight children) of the sum of the squared deviations of her predictions from the children's own responses.

To obtain pupils' descriptions of their teacher, a 12-item scale was developed with four items directed at each of the three kinds of teacher understanding we sought to measure. A mean score for each teacher was found for each item, based on the pupils' ratings on a four-point scale;

a total score was also obtained. The expectation of a general evaluative factor in these unforced ratings was confirmed by the intercorrelations and factor structure of the 12 items.

To secure differentiated descriptions of the teachers (possibly in terms of the three a priori categories) the 12 items were also presented to the pupils in all 66 possible pairs so as to comprise a forced-choice rating scale. With an especially written program, the Illinois electronic computer (Illiac) computed the intercorrelations between items. The forced-choice item correlation matrix yielded three main factors, two of which bore close resemblance to our a priori cognitive and personal problem categories. The third factor seemed to relate to children's judgments of knowledgeability of the teacher.

Intercorrelation of the understanding measures and the pupil ratings revealed only one significant correlation. This was an r of .28 between teacher's accuracy in predicting interpupil preferences and her pupils' judgement that their teacher "knows which pupils you like best in this class." This accuracy score also correlated significantly (.33) with accuracy in predicting pupils' problems. In addition, it correlated .27 with the mean socioeconomic status of the class, indicating some relationship between transparency of interpupil preferences and pupils' "social class" status. The correlational sociometric accuracy score tended to relate to these variables in the same direction, but not to the same degree. Its correlation with the score for accuracy in predicting specific choices was .46.

No significant curvilinear relationships were found between the understanding measures and the pupils' ratings. Similarly, there was no relationship between the pattern of the teachers' three accuracy scores and the pattern of the mean ratings they received from pupils on the three main factors of the forced-choice rating scale. Second-order relationships—in which the correlations between an understanding score and a mean rating were significantly different in groups of teachers stratified by grade level, age, sex, or score on the Minnesota Teacher Attitude Inventory—were significant at the .05 level in several instances. These are considered merely suggestive at present.

The report details the development and characteristics of each instrument. Inferences are drawn as to fruitful next steps in this research area. Particularly needed are techniques to identify areas of understanding that are relevant to teacher behavior, and techniques for identifying relevant motivations and skills.

REFERENCES

1. Ausubel, D. P., Schiff, H. M., & Gasser, E. B. A preliminary study of developmental trends in socioempathy: accuracy of perception of own and others' sociometric status. *Child Develpm.*, 1952, **23**, 111-128.
2. Bell, G. B., & Hall, H. E. The relationship between leadership and empathy. *J. abnorm. soc. Psychol.*, 1954, **49**, 156-157.
3. Cook, W. W., Leeds, C. H., & Callis, R. *Minnesota teacher attitude inventory: Manual.* New York: Psychological Corp., 1951.
4. Cronbach, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, **16**, 297-334.
5. Crow, W. J. A methodological study of social perceptiveness. Unpublished thesis, Univer. of Colorado, 1954.
6. Della Piana, G. M., & Gage, N. L. Pupils' values and the validity of the Minnesota Teacher Attitude Inventory. *J. educ. Psychol.*, 1955, **46**, 167-178.
7. Dymond, Rosalind F., Hughes, Anne S., & Raabe, Virginia L. Measurable changes in

empathy with age. *J. consult. Psychol.*, 1952, **16**, 202-206.

8. GAGE, N. L., & EXLINE, R. V. Social perception and effectiveness in discussion groups. *Hum. Relat.*, 1953, **6**, 381-396.
9. GRONLUND, N. E. The accuracy of teachers' judgments concerning the sociometric status of sixth-grade pupils. *Sociometry Monogr.*, No. 25, 1951.
10. GUILFORD, J. P. *Fundamental statistics in psychology and education.* (2nd Ed.) New York: McGraw-Hill, 1950.
11. GUTTMAN, L. Multiple rectilinear prediction and the resolution into components. *Psychometrika*, 1940, **5**, 75-99.
12. GUTTMAN, L. A basis for analysing test-retest reliability. *Psychometrika*, 1945, **10**, 255-282.
13. HEIDER, F. Attitudes and cognitive organization. *J. Psychol.*, 1946, **21**, 107-112.
14. HOLZINGER, K. J., & HARMAN, H. H. *Factor analysis.* Chicago: Univer. of Chicago Press, 1941.
15. HORST, P. A generalized expression for the reliability of measures. *Psychometrika*, 1949, **14**, 21-31.
16. JONES, L. V., & FISKE, D. W. Models for testing the significance of combined results. *Psychol. Bull.*, 1953, **50**, 375-382.
17. KELLEY, IDA B., & PERKINS, K. J. *How I teach: Purdue teacher's examination.* Minneapolis: Educational Test Bureau, 1942.
18. LEEDS, C. H. A scale for measuring teacher-pupil attitudes and teacher-pupil rapport. *Psychol. Monogr.*, 1950, **64**, No. 6 (Whole No. 312).
19. LORGE, I., & KRUGLOV, LORRAINE. The improvement of estimates of test difficulty. *Educ. psychol. Measmt.*, 1953, **13**, 34-46.
20. NEUHAUS, J. O., & WRIGLEY, C. F. The quartimax method: an analytic approach to orthogonal simple structure. *Brit. J. statist. Psychol.*, 1954, **7**, 81-91.
21. RABINOWITZ, W. The fakability of the Minnesota Teacher Attitude Inventory. *Educ. psychol. Measmt.*, 1954, **14**, 657-664.
22. SEARS, R. R. A theoretical framework for personality and social behavior. *Amer. Psychologist*, 1951, **6**, 476-483.
23. STONE, G. C., LEAVITT, G. S., & GAGE, N. L. Generality of accuracy in perceiving standard persons. Champaign-Urbana: Bureau of Educational Research, Univer. of Illinois, 1954. Rep. No. 1, Studies in the Generality and Behavioral Correlates of Social Perception. (Mimeographed)
24. TAFT, R. Some correlates of the ability to make accurate social judgments. Unpublished doctor's dissertation, Univer. of California, Berkeley, 1950.

(Accepted for publication June 16, 1955)